Hadamard Conjugation for the Kimura 3ST Model: Combinatorial Proof using Pathsets

Michael D. Hendy* Sagi Snir†

February 4, 2008

Abstract In most stochastic models of molecular sequence evolution the probability of each possible pattern of homologous characters at a site is estimated numerically. However in the case of Kimura's three-substitution-types (K3ST) model, these probabilities can be expressed analytically by Hadamard conjugation as a function of the phylogeny T and the substitution probabilities on each edge of T, together with an analytic inverse function. In this paper we produce a direct proof of these results, using pathset distances which generalise pairwise distances between sequences. This interpretation allows us to apply Hadamard conjugation to a number of topical problems in the mathematical analysis of sequence evolution.

*Allan Wilson Centre for Molecular Ecology and Evolution, Massey University, Palmerston North, New Zealand. m.hendy@massey.ac.nz

†**Corresponding author.** Mathematics dept. University of California, Berkeley, CA 94720, USA. ssagi@math.berkeley.edu

Key words : Hadamard conjugation, K3ST model, pathsets, phylogenetic trees, phylogenetic invariants.

1 Introduction

Hadamard conjugation is an analytic formulation of the relationship between the probabilities of expected site patterns of nucleotides for a set of homologous nucleotide sequences and the parameters of some simple models of sequence evolution on a proposed phylogeny T. An important application of these relations is to give a theoretical tool to analyse properties of phylogenetic inference, such as the methods of maximum likelihood and maximum parsimony, as well as being a tool for generating simulated data, and determining phylogenetic invariants. Hadamard conjugation can also be used as directly as for phylogenetic inference, inferring either trees with the Closest Tree algorithm [7, 18] or networks using Spectronet [11].

Hadamard conjugation was first introduced in 1989 [6, 8] to analyse two-state character sequences evolving under the Neyman model [15]. Evans and Speed in 1993 [5] noted that Kimura's three substitution types (K3ST) model [14] for 4-state characters could be modelled by the Klein group $\mathbb{Z}_2 \times \mathbb{Z}_2$. Noting this Székely et al [21, 22] extended the two-state analysis to a more general algebraic theory, where substitutions belonged to an arbitrary Abelian group. They then applied this to sequences evolving under the K3ST model. Current applications of Closest Tree and Spectronet [11] are usually applied to the $4-$state K3ST model or its derivatives, the K2ST and Jukes–Cantor models.

A pathset in a phylogenetic tree T, is a generalisation of the concept of a paths. This approach allows the concept of pairwise distances between sequences to be extended to distances connecting larger sets of taxa. It provides properties that can be related to other models, such as the molecular clock hypothesis. This has, for example, proved pivotal in allowing a simpler analytic expression of the likelihood function, as developed in [4], leading to an algebraic solution for the maximum likelihood points. It has also proved useful in identifying phylogenetic invariants [9], and to the introduction of projected spectra [23] which reduces both the variance in the parameter estimates, and the computational complexity of the Closest Tree algorithm [7]. Each of the above examples rely on some identities between the phylogenetic tree and the probabilities of obtaining sequences evolved under that tree. However, these identities were never *directly* proved. Here we provide for the first time, a direct proof for these identities. Effectively, this is an alternative proof of Hadamard conjugation for the K3ST model, where practical interpretations of the intermediate terms are developed, showing *directly* the relationships between the topology of T and the substitution probabilities across its edges. This is an important contribution that can serve in the burgeoning area of algebraic statistics in biology and phylogenetics, in particular (see e.g. [1, 2, 3, 16, 17, 20]).

We model the relationship of the differences of n sequences labeled $1, 2, \cdots, n$, from a reference sequence labeled 0. Because the models are reversible, the choice of reference sequence is arbitrary. The topology of T and the model parameters are presented in a sparse matrix Q_T of 2^n rows and columns, called the edge-length spectrum. The probabilities of each site pattern are presented in a similar sized matrix P_T called the sequence probability spectrum.

We also define a Hadamard matrix H_n of 2^n rows and columns, and show that the matrix products

$$H_nQ_TH_n, \quad H_nP_TH_n,$$

both relate to properties of path-sets. We prove the major result by interpreting corresponding components of each entry of these matrices.

In earlier representations [10, 19] the Hadamard conjugations for K3ST were presented as a conjugations of vectors of 4^n components by the Hadamard matrices H_{2n} of 4^n rows and columns. In the formulation presented here the vectors are replaced by matrices of 2^n rows and columns, which pre- and post-multiplied by H_n, a Hadamard matrix of the same order.

2 Kimura's 3ST model

Kimura's [14] three substitution types model (K3ST) specified independent rates, α, β and γ, for each of three substitution types between the RNA or DNA nucleotides. Here we will refer to these substitutions as:

t_α: the substitutions `A` ↔ `G, U(T)` ↔ `C` (transitions);

t_β: the substitutions `A` ↔ `U(T), G` ↔ `C` (transversions type β);

t_γ: the substitutions `A` ↔ `C, U(T)` ↔ `G` (transversions type γ).

By including the identity t_ϵ, we find the set of substitutions

$$\mathcal{T} = \{t_\epsilon, t_\alpha, t_\beta, t_\gamma\}$$

is a group under composition, which acts on the nucleotide set $\{\mathtt{A}, \mathtt{C}, \mathtt{G}, \mathtt{T(U)}\}$.

Observation 1 $(\mathcal{T}, \circ)$ *is isomorphic to the Klein* $4-$*group,* $(\mathbb{Z}_2 \times \mathbb{Z}_2, +_2)$.

□

Kimura modelled the expected differences between two sequences separated by time t. With the three specified rates, the expected numbers of substitutions of each type are therefore

$$q(\alpha) = \alpha t, \quad q(\beta) = \beta t, \quad q(\gamma) = \gamma t.$$

The number of substitutions of each type observed between homologous nucleotides of the two sequences can be used to estimate the probabilities $p(\alpha)$, $p(\beta)$, and $p(\gamma)$ of each type occurring. By setting $\beta = \gamma$, or $\alpha = \beta = \gamma$, this model projects to Kimura's better known two substitution type model [13], or to the simple Jukes/Cantor model [12].

Kimura derived expressions for the expected numbers as functions of the probabilities. These are equivalent to the standard expression of the rate matrix R derived from the stochastic matrix M, over time t,

$$M = \exp(Rt), \tag{1}$$

where, with $K = q(\alpha) + q(\beta) + q(\gamma)$ being the total number of substitutions,

$$M = \begin{bmatrix} p(\epsilon) & p(\alpha) & p(\beta) & p(\gamma) \\ p(\alpha) & p(\epsilon) & p(\gamma) & p(\beta) \\ p(\beta) & p(\gamma) & p(\epsilon) & p(\alpha) \\ p(\gamma) & p(\beta) & p(\alpha) & p(\epsilon) \end{bmatrix}, \quad Rt = \begin{bmatrix} -K & q(\alpha) & q(\beta) & q(\gamma) \\ q(\alpha) & -K & q(\gamma) & q(\beta) \\ q(\beta) & q(\gamma) & -K & q(\alpha) \\ q(\gamma) & q(\beta) & q(\alpha) & -K \end{bmatrix}.$$

Let H_2 be the 4×4 Hadamard matrix

$$H_2 = \begin{bmatrix} 1 & 1 & 1 & 1 \\ 1 & -1 & 1 & -1 \\ 1 & 1 & -1 & -1 \\ 1 & -1 & -1 & 1 \end{bmatrix}.$$

Observation 2 *H_2 diagonalises both M and Rt. In particular*

$$H_2^{-1} M H_2 = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 - 2p(\alpha) - 2p(\gamma) & 0 & 0 \\ 0 & 0 & 1 - 2p(\beta) - 2p(\gamma) & 0 \\ 0 & 0 & 0 & 1 - 2p(\alpha) - 2p(\beta) \end{bmatrix},$$

and

$$H_2^{-1} Rt H_2 = -2 \begin{bmatrix} 0 & 0 & 0 & 0 \\ 0 & q(\alpha) + q(\gamma) & 0 & 0 \\ 0 & 0 & q(\beta) + q(\gamma) & 0 \\ 0 & 0 & 0 & q(\alpha) + q(\beta) \end{bmatrix}.$$

□

Hence from equation 1 we find

$$\begin{aligned}
1-2(p(\alpha)+p(\gamma)) = p(\epsilon)-p(\alpha)+p(\beta)-p(\gamma) &= \mathrm{e}^{-2(q(\alpha)+q(\gamma)} = \mathrm{e}^{-K-q(\alpha)+q(\beta)-q(\gamma)}, \\
1-2(p(\beta)+p(\gamma)) = p(\epsilon)+p(\alpha)-p(\beta)-p(\gamma) &= \mathrm{e}^{-2(q(\beta)+q(\gamma)} = \mathrm{e}^{-K+q(\alpha)-q(\beta)-q(\gamma)}, \\
1-2(p(\alpha)+p(\beta)) = p(\epsilon)-p(\alpha)-p(\beta)+p(\gamma) &= \mathrm{e}^{-2(q(\alpha)+q(\beta)} = \mathrm{e}^{-K-q(\alpha)-q(\beta)+q(\gamma)},
\end{aligned}$$

which can be succinctly expressed as

$$H_1^{-1}PH_1 = \mathrm{Exp}(H_1^{-1}QH_1), \tag{2}$$

where

$$H_1 = \begin{bmatrix} 1 & 1 \\ 1 & -1 \end{bmatrix}, \quad P = \begin{bmatrix} p(\epsilon) & p(\alpha) \\ p(\beta) & p(\gamma) \end{bmatrix}, \quad Q = \begin{bmatrix} -K & q(\alpha) \\ q(\beta) & q(\gamma) \end{bmatrix},$$

and Exp is the exponential function to each entry of the matrix. Equation 2 can be inverted (provided the arguments of ln are all positive) to give

$$H_1^{-1}QH_1 = \mathrm{Ln}(H_1^{-1}PH_1), \tag{3}$$

where Ln is the natural logarithm applied to each component of the matrix.

The invertibility of equations 2 and 3 mean that provided the parameters are in the valid ranges, the model could be specified by the three probabilities $p(\alpha)$, $p(\beta)$ and $p(\gamma)$, or by the three parameters $q(\alpha)$, $q(\beta)$ and $q(\gamma)$. Indeed, when we do this, we do not need to rely on a rate/time specification and a Poisson process of substitution.

3 Substitutions across the edges of a tree

Let T be a tree (phylogeny) with leaf set $L(T) = \{0, 1, \ldots, n\}$, and edge set $E(T)$. We can postulate three independent Kimura probability parameters $p_e(\alpha)$, $p_e(\beta)$ and $p_e(\gamma)$ for each edge $e \in E(T)$ and a transition matrix

$$M_e = \begin{bmatrix} p_e(\epsilon) & p_e(\alpha) & p_e(\beta) & p_e(\gamma) \\ p_e(\alpha) & p_e(\epsilon) & p_e(\gamma) & p_e(\beta) \\ p_e(\beta) & p_e(\gamma) & p_e(\epsilon) & p_e(\alpha) \\ p_e(\gamma) & p_e(\beta) & p_e(\alpha) & p_e(\epsilon) \end{bmatrix}.$$

Suppose we assign nucleotides to each vertex of T according to a model parameterised by these probabilities for each edge $e \in E(T)$. The matrices M_e for each $e \in E(T)$ are all diagonalised by H_2, and hence commute, so for any subset $W \subseteq E(T)$ of edges we can define

$$M_W = \prod_{e \in W} M_e, \tag{4}$$

the transition matrix representing the probabilities of change concatenated across the edges in W.

We observe

$$\begin{aligned} H_2^{-1} M_W H_2 &= H_2^{-1} \left(\prod_{e \in W} M_e \right) H_2 \\ &= \prod_{e \in W} H_2^{-1} M_e H_2, \end{aligned}$$

is a diagonal matrix whose entries are the products of the corresponding eigenvalues of the factor matrices M_e.

We can define corresponding 2 × 2 matrices

$$P_W = \begin{bmatrix} p_W(\epsilon) & p_W(\alpha) \\ p_W(\beta) & p_W(\gamma) \end{bmatrix}, \quad Q_W = \begin{bmatrix} -K_W & q_W(\alpha) \\ q_W(\beta) & q_W(\gamma) \end{bmatrix},$$

writing P_e for $P_{\{e\}}$, etc. Hence, from equation 4, the entries $q_W(\alpha)$, $q_W(\beta)$ and $q_W(\gamma)$ of Q_W are linear functions of the logarithms of these eigenvalues, and we find

Observation 3

$$Q_W = \sum_{e \in W} Q_e.$$

□

Because of this linearity, we define $q_e(\alpha)$, $q_e(\beta)$ and $q_e(\gamma)$, to be the three **edge-length** parameters, for each edge e, and can specify our model by the $3|E(T)|$ independent parameters

$$q_e(\theta): \quad \theta \in \{\alpha, \beta, \gamma\}; e \in E(T).$$

The deletion of an edge $e \in E(T)$ induces two subtrees, whose leaf label sets partition $[n]_0$ into two subsets. We choose that subset $A \in [n]$ (the subset not containing 0) to index e as e_A. We incorporate the edge-length parameters into three vectors $\mathbf{q}_\alpha$, $\mathbf{q}_\beta$ and $\mathbf{q}_\gamma$ indexed by the 2^n subsets of $[n]$, where for $A \subseteq [n]$

$$(\mathbf{q}_\alpha)_A = \begin{cases} q_{e_A}(\alpha) & \text{if } e_A \in E(T), \\ -\sum_{e_B \in E(T)} q_{e_B} & \text{if } A = \emptyset, \\ 0 & \text{else,} \end{cases}$$

with similar structures for $\mathbf{q}_\beta$ and $\mathbf{q}_\gamma$. The entries in these vectors are ordered by the subsets of $[n]$ listed lexicographically: $\emptyset$, $\{1\}$, $\{2\}$, $\{1,2\}$, $\{3\}$, $\{1,3\}$, $\{2,3\}$, $\{1,2,3\}$, $\{4\}$, $\cdots$, etc.

We will also find it convenient to gather these three vectors into a $2^n \times 2^n$ matrix

$$Q_T = [q_{A,B}]_{A,B \subseteq [n]},$$

where

$$q_{A,B} = \begin{cases} q_{e_A}(\alpha) & \text{if } e_A \in E(T), B = \emptyset, \\ q_{e_B}(\beta) & \text{if } A = \emptyset, e_B \in E(T), \\ q_{e_A}(\gamma) & \text{if } A = B, e_A \in E(T), \\ -K_T & \text{if } A = B = \emptyset, \\ 0 & \text{else,} \end{cases}$$

and

$$K_T = \sum_{e \in E(T)} (q_e(\alpha) + q_e(\beta) + q_e(\gamma)) = \sum_{e \in E(T)} K_e.$$

Thus the leading column of Q_T is $\mathbf{q}_\alpha$, the leading row is $\mathbf{q}_\beta$, and the leading column is $\mathbf{q}_\gamma$, all other entries are 0, apart from the leading entry which is $-K_T$ (hence the sum of all entries of Q_T is 0). Q_T is referred to as the **edge length spectrum** for T. The positive entries of this spectrum identify the edges of T.

If we propose a sequence of nucleotides at leaf 0, then we can generate homologous sequences at each of the other leaves under this model. A common position in each of these sequences is called a site. If in an instance of such sequences, the character states at leaves $0, 1, \ldots, n$ are $\chi(0), \chi(1), \ldots, \chi(n)$, which partitions $[n]$ into the subsets

$$S_\theta = \{i \in [n] \colon t_\theta(\chi(0)) = \chi(i)\}, \text{ for } \theta \in \{\epsilon, \alpha, \beta, \gamma\}.$$

Thus for example $S_\emptyset$ is the set of leaves of $[n]$ with the same state as at 0. We index a site pattern by (A, B), the pair of subsets of $[n]$, where

$$A = S_\alpha \cup S_\gamma, \quad B = S_\beta \cup S_\gamma,$$

$$\mathbf{q}_\alpha = \begin{bmatrix} -K(\alpha) \\ q_1(\alpha) \\ q_2(\alpha) \\ 0 \\ q_3(\alpha) \\ q_{13}(\alpha) \\ 0 \\ q_{123}(\alpha) \end{bmatrix}, \mathbf{q}(\beta) = \begin{bmatrix} -K(\beta) \\ q_1(\beta) \\ q_2(\beta) \\ 0 \\ q_3(\beta) \\ q_{13}(\beta) \\ 0 \\ q_{123}(\beta) \end{bmatrix}, \mathbf{q}_\gamma = \begin{bmatrix} -K(\gamma) \\ q_1(\gamma) \\ q_2(\gamma) \\ 0 \\ q_3(\gamma) \\ q_{13}(\gamma) \\ 0 \\ q_{123}(\gamma) \end{bmatrix},$$

$$Q_T = \begin{bmatrix} -K & q_1(\alpha) & q_2(\alpha) & 0 & q_3(\alpha) & q_{13}(\alpha) & 0 & q_{123}(\alpha) \\ q_1(\beta) & q_1(\gamma) & . & . & . & . & . & . \\ q_2(\beta) & . & q_2(\gamma) & . & . & . & . & . \\ 0 & . & . & 0 & . & . & . & . \\ q_3(\beta) & . & . & . & q_3(\gamma) & . & . & . \\ q_{13}(\beta) & . & . & . & . & q_{13}(\gamma) & . & . \\ 0 & . & . & . & . & . & 0 & . \\ q_{123}(\beta) & . & . & . & . & . & . & q_{123}(\gamma) \end{bmatrix},$$

Figure 1: *Example edge length spectra for the tree T_{13} on $n+1 = 4$ taxa illustrated in figure 1. Corresponding components of the vectors $\mathbf{q}_\alpha$, $\mathbf{q}_\beta$, $\mathbf{q}_\gamma$, give the three edge lengths parameters for the corresponding edge. The value "0" value indicates that there is no corresponding edge in T. These vectors are placed in the leading row, column and main diagonal of the matrix Q. This means that for $A, B \subseteq \{1,2,3\}$, $Q_{\emptyset,B} = q_B(\alpha)$, $Q_{A,\emptyset} = q_A(\beta)$, $Q_{A,A} = q_A(\gamma)$, and for all other entries $Q_{A,B} = 0$, except the first entry $Q_{\emptyset,\emptyset} = -K$, where $K = K(\alpha) + K(\beta) + K(\gamma)$. The entries indicated by "." are all zero, these are zero for every tree. The entries indicated by "0" are zero for this tree T, but for different trees can be non-zero. The non-zero entries*

noting that the partition can be recovered from (A, B). In particular

$$S_\gamma = A \cap B,\ S_\alpha = A - S_\gamma,\ S_\beta = B - S_\gamma,\ S_\epsilon = [n] - (A \cup B).$$

We will show that the probability $p_{A,B}$ of obtaining the site pattern (A, B), for each $A, B \in [n]$, is a function of the edge length parameters.

We now define another $2^n \times 2^n$ matrix P_T, the **sequence probability spectrum**, with rows and columns indexed by the subsets of $[n]$, where

$$P_T = [p_{A,B}]_{A,B \subseteq [n]},$$

where p_{AB} is the probability of obtaining the site pattern (A, B).

4 Hadamard matrices and Path-sets

We define recursively the family $\{H_n : n \in \mathbb{Z}\}$, (known as Sylvester matrices), where for $n \geq 2$

$$H_n = H_1 \otimes H_{n-1} = \begin{bmatrix} H_{n-1} & H_{n-1} \\ H_{n-1} & -H_{n-1} \end{bmatrix}$$

is a symmetric Hadamard matrix of order 2^n, with H_1 and H_2 as previously defined. It is easily seen that $H_n^{-1} = 2^{-n} H_n$.

It is known [19] that if we index the rows and columns of H_n lexicographically by the subsets of $[n]$ that:

Observation 4

$$[H_n]_{A,B} = h(A, B) = (-1)^{|A \cap B|}.$$

□

Let $\Pi_{i,j}$ be the set of edges in the path in T connecting leaves i and j, $(i,j \in \{0,1,\ldots,n\})$ the entries of the transition matrix $M_{\Pi_{i,j}}$ represent the probabilities of observing the corresponding differences between the nucleotides at leaves i and j. We see further that

$$M_{\Pi_{i,j}} = \prod_{e_A \in \Pi_{i,j}} M_{e_A}.$$

Because each edge e_A in $E(T)$ separates vertices i from j, these edges are precisely those for which $A \cap \{i,j\}$ contains one, but not both elements. Hence we see

$$\Pi_{i,j} = \{e_A \in E(T) \colon h(A, \{i,j\}) = -1\}.$$

We generalise this, for any $C \subseteq [n]$, finding it useful to consider the collection of edges

$$\Pi_C = \{e_A \in E(T) \colon h(A, C) = -1\}.$$

Observation 5 *In [19] it is shown that:*

for $|C| \equiv 0(mod\ 2)$, Π_C is a set of $|C|/2$ edge-disjoint paths, whose endpoints are the leaves in C;

for $|C| \equiv 1(mod\ 2)$, Π_C is a set of $(|C|+1)/2$ edge-disjoint paths, whose endpoints are the leaves in $C \cup \{0\}$.

□

Π_C is called a **path-set**. In particular $\Pi_{\{i\}} = \Pi_{0,i}$ and $\Pi_{\{i,j\}} = \Pi_{i,j}$ comprise single paths, and $\Pi_\emptyset = \emptyset$. We find the set of pathsets is a group (under symmetric difference) isomorphic to $\mathbb{Z}_2^n$.

The sum of edge lengths on a path connecting two leaves can naturally be thought of as the distance between the leaves. We extend this distance concept, for each substitution type $\theta \in \{\alpha, \beta, \gamma\}$ to sets of paths, to define the **path-set distance**

$$d_{\Pi_C}(\theta) = \sum_{e_A \in \Pi_C} q_A(\theta),$$

so that

$$\begin{aligned}
\sum_{A \subset [n]} h(A,C) q_A(\theta) &= q_\emptyset + \sum_{e_A \in E(T)} h(A,C) q_A(\theta) \\
&= \sum_{e_A \in E(T)} (-1 + h(A,C)) q_A(\theta) \\
&= -2 \sum_{h(A,C)=-1} q_A(\theta) \\
&= -2 d_{\Pi_C}(\theta). \qquad (5)
\end{aligned}$$

Suppose each vertex v of T is assigned a character state $\chi(v)$, then for each edge $e = (u,v) \in E(T)$ there is a transformation t_{θ_e} such that $t_{\theta_e}(\chi(u)) = \chi(v)$. We can write $t_{\theta_e} = (\chi(u))^{-1}\chi(v) = \chi(u)\chi(v)$, as the transformations are Boolean. For the path $\Pi_{i,j}$ connecting leaves i and j we find

$$\prod_{e \in \Pi_{i,j}} t_{\theta_e} = \prod_{e=(u,v) \in E(T)} \chi(u)\chi(v) = \chi(i)\chi(j),$$

as the products at each internal vertex cancel. Further, for any $C \subseteq [n]$ let $C_0 = C \cup \{0\}$ if $|C|$ is odd, $C_0 = C$ otherwise, then

$$\prod_{e \in \Pi_C} \theta_e = \prod_{i \in C_0} \chi(i). \qquad (6)$$

Suppose $\prod_{e \in \Pi_C} \theta_e \in \{\epsilon, \alpha\}$, then the number of factors $\chi(i) \in \{\beta, \gamma\}$ in equation 6 must be even, hence with $B = S_\beta \cup \gamma$, $h(B, C_0) = h(B, C) = 1$. Similarly if $\prod_{e \in \Pi_C} \theta_e \in \{\beta, \gamma\}$, then the number of factors $\chi(i) \in \{\beta, \gamma\}$ in equation 6 must be odd, so $h(B, C_0) = h(B, C) = -1$. Generalising this we obtain

Observation 6 *If each vertex v of T is assigned character state $\chi(v)$ with $A = S_\alpha \cup S_\gamma, B = S_\beta \cup \gamma \subseteq [n]$, and for any $C \subseteq [n]$ then for any $C \subseteq \{0, 1, \cdots, n\}$ with an even number of elements, let $\chi(C) = \prod_{i \in C} \chi_i$ then*

$$\chi(C) \in \{\epsilon, \alpha\} \Leftrightarrow h(B, C) = 1, \quad \chi(C) \in \{\epsilon, \beta\} \Leftrightarrow h(A, C) = -1. \tag{7}$$

5 Hadamard Conjugation

Q_T is the matrix containing the edge weight parameters across T. P_T is the matrix of probabilities of patterns at the leaves of T. The link between these are the matrix products $H_n P_T H_n$ and $H_n Q_T H_n$ which both relate to pathset properties. These enable to state our major result

Theorem 7

$$P_T = H_n^{-1}(\text{Exp}(H_n^{-1} Q_T H_n))H_n, \tag{8}$$

which provided the arguments of the logarithm are positive, is invertible to give

$$Q_T = H_n^{-1}(\text{Ln}(H_n P_T H_n))H_n^{-1}, \tag{9}$$

Proof

The proof of this theorem is based on interpreting the corresponding components, for $A, B \subseteq$

$[n]$,

$$[H_nP_TH_n]_{A,B} \text{ and } [H_nQ_TH_n]_{A,B}\,.$$

As the only nonzero entries in Q_T are $Q_{\emptyset,\emptyset}$ and $Q_{C,\emptyset}, Q_{\emptyset,C}, Q_{C,C}\colon e_C \in E(T)$, we find

$$\begin{aligned}
[H_nQ_TH_n]_{A,B} &= \sum_{A',B'\subseteq[n]} h(A,A')h(B,B')Q_{A'B'} \\
&= Q_{\emptyset,\emptyset} + \sum_{e_C\in E(T)} (h(A,C)Q_{C,\emptyset} + h(B,C)Q_{\emptyset,C} + h(A,C)h(B,C)Q_{C,C} \\
&= \sum_{e_C\in E(T)} ((h(A,C)-1)Q_{C,\emptyset} + (h(B,C)-1)Q_{\emptyset,C} + (h(A,C)h(B,C)-1)Q_{C,C}) \\
&= \sum_{e_C\in E(T)} ((h(A,C)-1)q_{e_C}(\beta) + (h(B,C)-1)q_{e_C}(\alpha) + (h(A,C)h(B,C)-1)q_{e_C}(\gamma)) \\
&= -2\sum_{e_C\in\Pi_A} q_{e_C}(\beta) - 2\sum_{e_C\in\Pi_B} q_{e_C}(\alpha) - 2\sum_{e_C\in\Pi_A\Delta\Pi_B} q_{e_C}(\gamma) \\
&= -2\left(d_{\Pi_A}(\beta) + d_{\Pi_B}(\alpha) + d_{\Pi_A\Delta\Pi_B}(\gamma)\right).
\end{aligned}$$

We can partition $\Pi_A \cup \Pi_B$ into three parts,

$$U = \Pi_A - \Pi_B, \quad V = \Pi_B - \Pi_A, \quad W = \Pi_A \cap \Pi_B,$$

and likewise split the pathset distances into components $d_U(\theta) = \sum_{e\in U} q_e(\theta)$, etc. Thus

$$[H_nQ_TH_n]_{A,B} = -2\left(d_U(\beta) + d_U(\gamma) + d_V(\alpha) + d_V(\gamma) + d_W(\alpha) + d_W(\beta)\right),$$

and

$$[\mathrm{Exp}(H_nQ_TH_n)]_{A,B} = \mathrm{e}^{-2(d_U(\beta)+d_U(\gamma))}\mathrm{e}^{-2(d_V(\alpha)+d_V(\gamma))}\mathrm{e}^{-2(d_W(\alpha)+d_W(\beta))}. \tag{10}$$

Now, by equation 2,

$$\begin{aligned}
\mathrm{e}^{-2(d_U(\beta)+d_U(\gamma))} &= p_U(\epsilon) + p_U(\alpha) - p_U(\beta) - p_U(\gamma), \\
\mathrm{e}^{-2(d_V(\alpha)+d_V(\gamma))} &= p_V(\epsilon) - p_V(\alpha) + p_V(\beta) - p_V(\gamma), \\
\mathrm{e}^{-2(d_W(\alpha)+d_W(\beta))} &= p_W(\epsilon) - p_W(\alpha) - p_W(\beta) + p_W(\gamma).
\end{aligned}$$

Hence equation 10 becomes

$$
\begin{aligned}
[\mathrm{Exp}(H_nQ_TH_n)]_{A,B} &= (p_U(\epsilon)+p_U(\alpha)-p_U(\beta)-p_U(\gamma)) \\
&\quad \times(p_V(\epsilon)-p_V(\alpha)+p_V(\beta)-p_V(\gamma)) \qquad (11)\\
&\quad \times(p_W(\epsilon)-p_W(\alpha)-p_W(\beta)+p_W(\gamma)),
\end{aligned}
$$

which, when expanded, comprises the sum of 64 terms of the form

$$\pm p_U(\theta)p_V(\phi)p_W(\psi), \quad \theta,\phi,\psi\in\{\epsilon,\alpha,\beta,\gamma\}.$$

Now consider the joint probability $Pr[\Pi_A\colon\alpha;\Pi_B\colon\beta]$ that the product of substitutions across the edges of Π_A is t_α and the product across the edges of Π_B is t_β. This event is attained by the combinations of t_θ across U, t_ϕ across V and t_ψ across W such that

$$t_\theta t_\psi = t_\alpha \text{ and } t_\phi t_\psi = t_\beta,$$

which is attained with $t_\theta = t_\alpha t_\psi$ and $t_\phi = t_\beta t_\psi$, for each $\psi\in\{\epsilon,\alpha,\beta,\gamma\}$, and hence with probability

$$Pr[\Pi_A\colon\alpha;\Pi_B\colon\beta] = p_U(\alpha)p_V(\beta)p_W(\epsilon)+p_U(\epsilon)p_V(\gamma)p_W(\alpha)+p_U(\gamma)p_V(\epsilon)p_W(\beta)+p_U(\beta)p_V(\alpha)p_W(\gamma),$$

and these terms each occur with a + sign in equation 11.

Now we see the joint probability that the product of substitutions across Π_A is either ϵ or α and the product across Π_B is either ϵ or β is

$$Pr[\Pi_A\colon\epsilon,\alpha;\Pi_B\colon\epsilon,\beta] = Pr[\Pi_A\colon\epsilon;\Pi_B\colon\alpha]+Pr[\Pi_A\colon\epsilon;\Pi_B\colon\alpha]+Pr[\Pi_A\colon\epsilon;\Pi_B\colon\alpha]+Pr[\Pi_A\colon\epsilon;\Pi_B\colon\alpha],$$

and each summand appears with positive sign in equation 11. Similar examinations of the terms of equation 11 gives

$$[\mathrm{Exp}(H_nQ_TH_n)]_{A,B} = Pr[\Pi_A\colon\epsilon,\alpha;\Pi_B\colon\epsilon,\beta]+Pr[\Pi_A\colon\beta,\gamma;\Pi_B\colon\alpha,\gamma]$$

$$-Pr[\Pi_A\colon \epsilon,\alpha;\Pi_B\colon \alpha,\gamma] - Pr[\Pi_A\colon \beta,\gamma;\Pi_B\colon \epsilon,\beta]. \tag{12}$$

Let A_0 be the set of endpoints of Π_A, ($A_0 = A$ or $A \cup \{0\}$, whichever is of even order). Then

$$\prod_{e=(u,v)\in\Pi_A} \chi(u)\chi(v) = \prod_{i\in A_0} \chi(i),$$

as for each internal vertex w, $\chi(w)$ occurs twice in the product, which gives the identity. Hence $Pr[\Pi_A\colon \epsilon,\alpha;\Pi_B\colon \epsilon,\beta]$, the joint probability that the product of substitutions across Π_A is either ϵ or α and the product across Π_B is either ϵ or β, is the joint probability that the product of of states across the leaves of A_0 is ϵ or α and across the leaves of B_0 is ϵ or β. Thus by equation 7

$$Pr[\Pi_A\colon \epsilon,\alpha;\Pi_B\colon \epsilon,\beta] = \sum_{A',B'\subseteq[n]:h(A,A')=1,h(B,B')=1} P_{A'B'}.$$

Similarly we find

$$\begin{aligned}
Pr[\Pi_A\colon \beta,\gamma;\Pi_B\colon \alpha,\gamma] &= \sum_{A',B'\subseteq[n]:h(A,A')=-1,h(B,B')=-1} P_{A'B'},\\
Pr[\Pi_A\colon \epsilon,\alpha;\Pi_B\colon \alpha,\gamma] &= \sum_{A',B'\subseteq[n]:h(A,A')=1,h(B,B')=-1} P_{A'B'},\\
Pr[\Pi_A\colon \beta,\gamma;\Pi_B\colon \epsilon,\beta] &= \sum_{A',B'\subseteq[n]:h(A,A')=-1,h(B,B')=1} P_{A'B'}.
\end{aligned}$$

Thus from equation 12, combining these probabilities we find

$$[\mathrm{Exp}(H_nQ_TH_n)]_{A,B} = \sum_{A',B'\subseteq[n]} h(A,A')h(B,B')P_{A'B'}, \tag{13}$$

giving

$$\mathrm{Exp}(H_nQ_TH_n) = H_nP_TH_n, \tag{14}$$

from which equations 8 and 9 follow.

□

References

[1] Allman, E.S. and J.A. Rhodes, 2003. Phylogenetic invariants for the general Markov model of sequence mutation, Mathematical Biosciences: **186**, 113–144.

[2] Allman, E.S. and J.A. Rhodes, 2004. Quartets and parameter recovery for the general Markov model of sequence mutation, Applied Mathematics Research eXpress: **4**, 107–131.

[3] Allman, E.S. and J.A. Rhodes, 2004. Phylogenetic ideals and varieties for the general Markov model. (Preprint)

[4] Chor, B., M.D. Hendy and S. Snir 2005. Maximum Likelihood Jukes-Cantor Triplets: Analytic Solutions. *manuscript* (oai:arXiv.org:q-bio/0505054)

[5] Evans, S.N., and T.P. Speed, 1993. Invariants of some probability models used in phylogenetic inference. Ann Statist. **21**:355-377.

[6] Hendy, M.D., 1989. The relationship between simple evolutionary tree models and observable sequence data. Syst. Zool. **38**:310-321.

[7] Hendy, M.D. 1991 A combinatorial description of the closest tree algorithm for finding evolutionary trees, Disc. Math. **96**, 51-58.

[8] Hendy, M.D., and D. Penny 1989. A framework for the quantitative study of evolutionary trees. Syst. Zool. **38**:297-309.

[9] Hendy, M.D. and D.Penny 1996 Complete families of linear invariants for some stochastic models of sequence evolution with and without the molecular clock assumption, J. Comp. Biol., **3**: 19-31.

[10] Hendy, M.D., D. Penny and M.A. Steel 1994. A discrete Fourier analysis for evolutionary trees. Proc. Natl. Acad. Sci. USA. **91**:3339-3343.

[11] Huber, K.T., M. Langton, D. Penny, V. Moulton and M. Hendy, 2002. Spectronet: A package for computing spectra and median networks, App. Bioinf., **1**: 159-161.

[12] Jukes, T.H. and C.R. Cantor 1969. Evolution of protein molecules, in H.N Munro (Ed.), Mammalian Protein Metabolism III, Academic Press, New York.

[13] Kimura, M. 1980. A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences. J. Mol. Evol. **16**:111-120.

[14] Kimura, M. 1981. Estimation of evolutionary distances between homologous nucleotide sequences. Proc. Natl. Acad. Sci. USA. **78**:454-458.

[15] Neyman, J.l 1971. Molecular studies of evolution: a source of novel statistical problems, in S.S. Gupta, J. Yackel (Eds.), Statistical Decision Theory and Related Topics, Academic Press, New York.

[16] Pachter, L. and B. Sturmfels, 2005. Algebraic Statistics for Computational Biology, Cambridge University Press.

[17] Pachter, L. and B. Sturmfels, 2005. The Mathematics of Phylogenomics, (Submitted)

[18] Steel, M.A., M.D. Hendy, L.A. Székely and P.L. Erdös 1992. Spectral analysis and a closest tree method for genetic sequences, Appl. Math. Lett. **5**:63-67.

[19] Steel, M.A., M.D. Hendy and D.Penny, 1998. Reconstructing phylogenies from nucleotide pattern probabilities: A survey and some new results, Disc. Appl. Math., **88**: 367-396.

[20] Sturmfels, B. and S. Sullivant, 2005. Toric ideals of phylogenetic invariants, Journal of Computational Biology **12**, 204–228.

[21] Székely, L., P.L. Erdös, M.A. Steel, and D. Penny, 1993. A Fourier inversion formula for evolutionary trees. Applied Mathematics Letters, **6**: 13-17.

[22] Székely, L., M.A. Steel, and P.L. Erdös, 1993. Fourier calculus on evolutionary trees. Advances in Applied Mathematics, **14**: 200-216.

[23] Waddell P.J., and M.D. Hendy. 1997. Using phylogenetic invariants to enhance spectral analysis of nucleotide sequence data, Information and Mathematical Sciences Report Series B, #97/01, Massey University, New Zealand.